Data wrangling report

Analysis: we rate dogs

July 7, 2022
Prepared by: Oluwasola Aduewa
Data analytics ND

This report is to provide an update on the data wrangling project that is due on June 29. The main objective of this project is to gather, assess, and clean data provided by a Twitter user **@dog_rates.** They are popular for their twitter display name, **WerateDogs.** The said twitter handle had downloaded and sent a copy of their tweet archive to Udacity for student learning purposes. The tweet archive contains basic tweet data (tweet id, timestamp, text etc.) for 5000+ of their tweets between when they started and the 1st of August 2017. I present in more details the stages been through to get the master dataset ready.

Data Gathering

All dataset were downloaded programmatically. Two of these datasets are available on Udacity and they have been saved as **twitter_archive_enhanced.csv** and **image_predictions.tsv.** The additional data gathered from Twitter API which originally was saved as **tweet_archive.json** was processed to fit into Pandas Data Frame where it was later saved as complementary_tweet_archive.csv.

Assessing Data:

The datasets were assessed both visually and by programmatic means. The following quality and tidiness issues were noted in the process:

Quality

- Some columns like in_reply_status_id, retweeted_status_id, etc. were redundant
- Some rating scores were wrongly matched from their texts
- Some dog names were wrongly matched from corresponding text
- Unrealistic rating scores were found for some records
- Some columns were not properly formatted
- Duplicate entries were found in some columns
- Some columns names were not descriptive enough
- Rating scores consist of outliers
- Some records contain empty entries

Tidiness issues

- Text column contain tweet URL, hashtags and texts which were unnecessary
- The stages of dogs which could have occupied a single column existed in their individual columns
- Varying predictions of animals were contained in the image prediction dataset rather than having true prediction for dogs only.

Data Cleaning:

Prior to cleaning the datasets, a copy of each dataset was made. Data cleaning was performed on the copies of the original. Tidiness issues were fixed before addressing the quality issues with the datasets. Some of the issues that were cleaned are as follows:

- Tweet URLs contained in the text column were extracted into a separate column whilst expunging them from original source
- Hashtags were completely removed from text column to occupy separate column
- Records of best dog predictions alongside their confidence level alongside key variables were filtered into a separate data frame named as *dog_prediction* **(5 columns, 1751 rows**)
- Dog stages were compressed into a single column
- Poorly formatted fields were converted to their appropriate formats
- Non-descriptive column names were renamed
- Redundant fields were dropped completely from the data
- Some dog names were rematched from their corresponding text to replace the wrongly matched entries
- Rating score entries were rematched from their corresponding texts to replace the ready-made rating columns
- Empty entries in hashtag and dog stage columns were filled with descriptive words that clearly suggest their absence

Storing Data

The cleaned versions of the three datasets were merged into a single master data frame. The master dataset was saved as twitter_archive_master.csv

Conclusion

The data wrangling procedure currently has no obvious obstacles or issues, except for the rating score outliers which were left as a result of the unique @WerateDogs system. Although not included in the master dataset, incorporating the ratio of the rating scores will be useful.